Comparative Analysis of K means Clustering Sequentially And Parallely

Kavya D S [1], Chaitra D Desai[2]

[1]M.tech, Computer Science and Engineering, REVA ITM, Bangalore, India
[2]REVA ITM, Bangalore, India

Abstract - *In Data mining one of the most widely studied topics is clustering or Cluster analysis. Clusters are formed by grouping input data sets in which it contains most similar elements and the process is called clustering. Clustering is been included in many fields like Pattern Recognition ,Bioinformatics, Image Analytics ,Machine Learning etc, as it is common technique used by statistical data analysis. Except the input data sets no extra knowledge or information will be given to clustering as in general it is unsupervised learning task. using this task many algorithm has been developed in many fields on clustering. In most of the clustering algorithms the task is computationally expensive due to its recursive or iterative procedures. In many application clustering plays a very importance role in solving problem. As their are many clustering algorithm to solve problems ,one of the commonly used clustering algorithm is K Means. Basically this algorithm works on distance measure ,as it is going to calculate the distance between the each data set and centroid, by distance calculation if that dataset is near to that centroid then that dataset belong to that cluster. This will be looped until their is no changes in any dataset. In this paper the performance evaluation is carried out with respect to time, using k means both sequentially and parallely on same datasets.*

Key words: Clustering, K means algorithm, Hadoop, Map reduce.

1. INTRODUCTION

Set of input patterns are grouped into disjoint clusters and this processing called clustering or cluster analysis. The processing is done in such a way that the similar input patterns or elements will be grouped into one cluster and dissimilar elements will be in another cluster. Data exploration technique is used by clustering in order to group the similar characteristics of object for their further processing. In many applications clustering plays a very important role including artificial intelligence, neural networks and statistics. In many fields like biochemistry, image/video processing, bio-infomatics clustering task has been exploited as it is a unsupervised learning task. Depending on the purpose of clustering or that properties many different kinds of algorithm has been developed on clustering, such as hierarchical, partitional, graph based etc. In-order to find optimal solutions globally or locally from the data sets which are high dimensional ,since the clustering task requires recursive procedure. The unique clustering solution can be rarely presented by real-life data but interpreting the cluster representation is very difficult. Hence on same set of data with different algorithms many experimentations should be performed. As these task are expensive with respect to computation and it is complex , as it is one of the issue in clustering algorithm[1].In many applications such as vector quantization, data mining, pattern classification[4], data compression ,as it arises a problems with respect to clustering, the notion is of what makes a good cluster, as it depends on many methods and that is been subjected to many criteria. since the decisions must be taken in order to form a good clustering, as their is no any best criteria which aims at good clustering, as clustering will be done in user needs.

Clustering can be done in two ways -1.Distance based clustering 2.conceptual clustering. In distance based ,the clustering will be done based on distance criteria. If more than one object have the similar measured distance then they belong to one cluster and rest of the objects will belong to other cluster depending on the distance. In conceptual clustering according to descriptive concepts the object will be grouped into clusters. Their are many clustering algorithms like K Means, Mixture of Gaussians, Fuzzy c-means etc.

2.KMEANS

Popular data clustering is K means as it is one of the unsupervised learning algorithm. The pre-specification should be made with respect to number of clusters in data in order to use K means algorithm. For many experimental application, K Means algorithm has been succeeded in providing a very good results in clustering ,for the given input values finding the number of clusters is basically analysis process which makes more complex by subjective

nature in order to decide whether it is good clustering[2]. Through the certain number of clusters, the input sets can be classified ,as the procedure is as simple and easy. Their will be many number of clusters depending on input sets, as for each cluster define K-centroids. The centroids will be placed in such a way that they be far away from each other. Each and every data points of a input set must belong to nearest centroid. Results from previous step must be considered, K-centroids of the cluster should be recalculated. New binding must be done between nearest new centroid and input sets once these K-new centroid are been formed. This should be continued until their is no new centroid can be created. Minimizing objective function is main aim of K Means.

Example : Let us Consider object consist of 4 chemicals and these object includes index and weight as two attribute

Object	Index	Weight
Chemical A	1	1
Chemical B	2	1
Chemical C	4	3
Chemical D	5	4

Table -1 : Object with index and weight

The attributes of each chemicals can be considered as co-ordinates.

1.Intially random values considered by taking chemicals A and B as centroids.
c1=(1,1) and c2=(2,1)
c1 and c2 are centroid co-ordinates.

2.By using Euclidean formula the distance between each object and centroid should be calculated.

$$\begin{matrix} A & B & C & D \\ 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{matrix}$$

c1=(1,1) Group 1
c2=(2,1) Group 2

Here A, B, C,D represent the objects and corresponding row represents the attributes. since the distance between object C and group 1 is calculated by

$\sqrt{(4-1)^2 + (3-1)^2} = 3.61$ for c1=(1,1)

$\sqrt{(4-2)^2 + (3-1)^2} = 2.83$ for c2=(2,1)

After calculating the distance between each and every object and centriod, can determine which object belong to which group.

$$\text{Distance} = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix}$$

$$\text{Group} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

Group 1
Group 2

with respect to group,1 indicates the object belong to particular group and 0 indicates it does not belong to that group.

3.Considering previous step, again K-centroids of the cluster must be recalculated. New binding must be done between nearest new centroid and input sets once these K-new centroid are been formed. This should be continued until their is no new centroid can be created.

The final distance and group matrix is given as

$$\text{Distance} = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix}$$

$$\text{Group} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Finally object A and B belongs to group 1 and object C and D belongs to group 2.

Algorithm

In K Means numbers of clusters are assumed to be fixed. To one of the n input patters (M1,....,Mn) let k prototypes (Z1,...., Zk) be initialized [5] .

Initialize K protypes (Z1,....,Zk) such that Zj=Mi
j belongs to (1,....,k), i belongs to (1,...,n)
with protype Zj will be associated with Each cluster Cj
Loop
for each input vector Mi , where l belongs to (1,..n)
do
with nearest Zj ,assign Mi to the cluster Cj
for each cluster Cj ,do
for all the samples presently update Zj to be centroid
Compute the Objective function
Until no longer cluster member changes.

3 SEQUENTIAL COMPARITIVE ANALYSIS

sequential execution means executing the instructions one after the other, since other instructions which has to be executed must be waiting in the queue. Here the input sets should be stored into the memory and execution will definitely depend on the size of the input. If the input size is large it takes time for executing or else it can be executed in certain time.

we are considering 20 Newsgroup dataset for sequential execution , as dataset consist of 20 sub-directory and each sub-directory has 1000 files with it. Initially the data should be converted into vectors form and also initial clusters has to be generated. Here Execution will takes place in two phases -iteration phase and cluster formation[table1]. In

this two phase we are getting execution time information with respect to Real time ,User time, system time.

Real - Time between start of execution and termination of process.

User - It is the amount of time spent with respect to CPU outside the Kernel *within* the processing user-mode.

Sys - It is the amount of time spent with respect to CPU within the process inside the kernel.

Input Datasets	Iteration Process	Cluster formation	Elapsed time
20 Newsgroup	41sec.109 ms	12sec.039ms	53sec.148ms

Table-2: Sequential Elapsed time

The elapsed time is calculated only by considering Real time

Elapsed Time = Iteration process+ cluster formation
= 41sec.109mseconds +12sec.039 ms
= 53sec.148 ms

The time taken by serial execution is 53sec and 148mseconds

4.OVERVIEW OF HADOOP

Apache software foundation developed a open-source framework called Hadoop[6].In the year 2003 ,white paper was published by google labs which describes the google architecture using GFS. Map reduce was published in December 2004,HDFS was a new technique was launched by Yahoo. In order to run their applications on clusters many companies like yahoo, face book, Amazon and Newyork Times has been successfully used Hadoop. It works on large datasets ,as the data is been stored and processed across cluster in distributed environment. Hadoop provides a combination of storage as well as processing , as it is one of the main advantage in Hadoop. Basically Hadoop works on commodity hardware as it is inexpensive and the storage is reliable. Hadoop hides the details of data distribution to processing nodes, consolidation of results after computation ,restarting failed nodes, parallel processing. Hadoop consist of two components mainly :HDFS and MapReduce. HDFS is used for storing the data and MapReduce is used to process those stored datasets.

HDFS

HDFS is specially designed file-system for storing huge datasets with cluster of commodity hardware and with streaming access pattern. HDFS consist of Name-node and Data-node[3].whereas Name-node will act as master and Data-node will act as slave. The daemon called job-tracker which runs on Name-node. As name-node manages and maintains the blocks on Data-node. Name-node has the Metadata, as it is the information that where the data has been stored in Data-node. The Name-node should be runned on high reliable hardware, since it is single point of failure, if Name-node down then none of the nodes works.

Data-node will act as a slave whereas the daemon called Task-tracker runs on Data-node. The services or process which runs in background is called Daemon. Data-node has the actual storage that whatever information that is present in Metadata.
Hadoop maintains multiple copies of data on Data-node in order to achieve fault tolerance. This process is called data replication[7]. Minimum 3 copies of data is been replicated. Basically all the operations in HDFS is carried out in terms of blocks. Block size is 64MB by default.

MAPREDUCE

It is a programming model as it is designed in such a way that it process the huge datasets which is stored in HDFS and generates the results across cluster in distributed environment[3]. Map reduce makes the developers to focus on data processing and it hides the parallel execution details. In-order to handle large web search applications, map reduce was proposed by google. This approach is effective programming for developing many applications like data mining, machine learning etc [8]. Map reduce consist of two processing functions called 1. Map 2. Reduce. The input datasets are been split into independent data chucks by map reduce job. these chucks will be given as input to map task. As map task process this input parallely and produces a intermediate results in the form of <key, Value> pair. This Intermediate results will be sorted and this will be given as input to the reduce phase. Reduce Phase aggregates them and produces the final output in the form of <key, Value> pair. Each and every node in map reduce consist of job tracker and Task tracker. Master will be job tracker and it resides in NameNode. slave will be TaskTracker and it resides in DataNode.

5 PARALLEL COMPARITIVE ANALYSIS

Hadoop is a very good framework in order to run the programs parallely ,as it is open source. Basically Hadoop works on commodity hardware as it is inexpensive and the storage is reliable. Hadoop installation provides different modes of cluster to be setup : 1.single node Hadoop cluster 2.multi node Hadoop cluster. To set up a Single node Hadoop cluster it provides a option in two different modes - standalone and pseudo-distributed mode.MapReduce allows the programmers to write mapper and reducer function. In this by default Hadoop provides 2 map and

1reduce slots in order to process the input. These slots will depend on the input, if the input file is large then 2 map slots will be used or else 1 map slot will be used. Initially input file or dataset must be stored in HDFS from the local machine.

Input set	Sequential Elapsed time	Parallel Elapsed Time
20 newsgroup	53sec.148msec	48 sec

Table -3: Overview of Jobs

As 20 Newsgroup dataset has already been discussed, Initially the data will be converted into vectors form and also initial clusters will be generated. while executing, the number of iterations must be specified, as per the specification the algorithm will run that many iterations. Hence that many mapreduce jobs takes place in sequence.

Input set	Job	Start time	Stop time	Elapsed time
20newsgroup	1	13:29:34	13:29:50	16 sec

Table-4: Elapsed Time with Job1

Elapsed time is calculated by difference between start and stop time of the process.
Elapsed Time =Start time - Stop time

Input set	Job	Start time	Stop time	Elapsed time
20newsgroup	2	13:29:59	13: 30:16	16 sec

Table-5: Elapsed Time with Job2

Input set	Job	Start time	Stop time	Elapsed time
20newsgroup	3	13:30:26	13: 30:42	16 sec

Table -6: Elapsed Time with Job3

The total elapsed time is calculated
Elapsed Time = Job 1 + job 2 + job 3
= 16sec +16 sec+16sec
= 48 sec

Input Datasets	Job 1	Job 2	Job 3	Elapsed Time
20 Newsgroups	16 sec	16 sec	16 sec	48 seconds

Table -7*:* Total Elapsed Time

Time taken by parallel execution 48 seconds.

6 COMPARASION BETWEEN SEQUENTIAL AND PARALLEL PROCESSES

Here we are comparing two execution process 1. sequential 2. parallel. Sequential is a process which runs the instruction step by step whereas parallel execution takes place in distributed manner. The comparison is done with respect to this two different processes by using same datasets and also same algorithm

Table-8: Elapsed times between processes

Sequential time = 53 sec.148msec
Parallel time = 48 sec
Difference = Sequential – parallel
= 53sec.148msec - 48 sec

Input set	Iterations	Jobs 1	Job 2	Job 3
20newsgroup	3	Map-2	Map -2	Map-2
		Reduce -1	Reduce -1	Reduce-1

= 5sec.148msec

From the above table, it clearly indicates that the time taken by sequential execution is higher than parallel execution. Their is a difference of 5seconds and 148milliseconds between sequential and parallel execution. By these results we can conclude that executing parallel is much more better than executing sequentially.

7 EXPERIMENTAL RESULTS

To evaluate the performance of parallel and sequential process we carried out some experiments by using case study Kmeans. Performance is been evaluated in both sequential and parallelly by considering different iterations and calculating the elapsed time in these iterations process. The Elapsed time which gives execution time(start and stop of process).This experiment is been carried out in both sequential and parallel and the results are been tabulated in tables.

For both sequential and parallel execution the input is 20_newsgroups dataset with fixed number of clusters=20

Iterations	Iteration Process	Cluster Formation	Elapsed Time
i=3	41s.109 ms	12s.039ms	53s.148ms
i=4	48s.426ms	21sec.724ms	1m.9s.1150ms
i=5	1m.2s.920ms	20sec.950ms	1m.22s.1870ms

Table -9: Sequential Execution with different Iterations

Iterations	Job1	Job2	Job3	Job4	Job5	Elapsed Time
i=3	16sec	16sec	16sec	_	_	48sec
i=4	15sec	16sec	16sec	15sec	_	1m.2s
i=5	15sec	15sec	15sec	15sec	16s	1m.16s

Table -10: Parallel Execution with different Iterations

Iterations	Sequential elapsed time in sec	Parallel-elapsed time in sec
i=3	53	48
i=4	70	62
i=5	84	76

Table-11: Elapsed time in seconds

The above tabulated results are been plotted in graph.

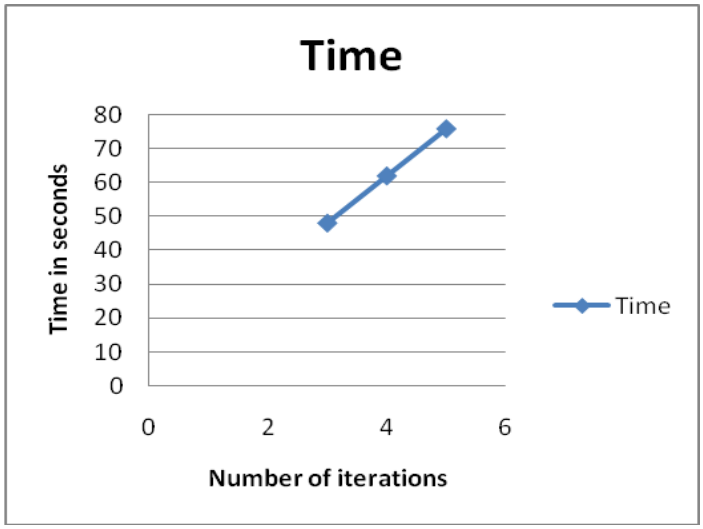

Chart-1: Sequential

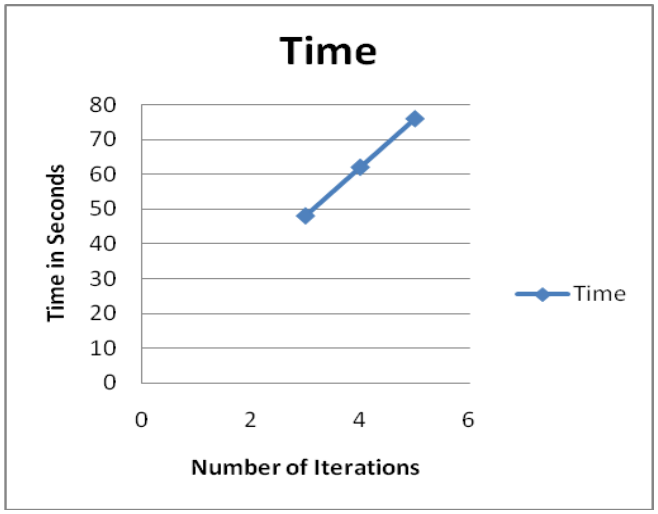

Chart-2: Parallel

From the above results we can tell that parallel execution is much efficient than sequential, as it saves the computation time.

8. CONCLUSION

In this paper we studied about clustering, how the clustering is formed, KMeans clustering algorithm and also comparison analysis on sequential and parallel form by using same datasets and same algorithm. By this we can conclude that parallel execution is efficient and processing the data is much faster than sequential.

9.FUTURE WORK

Depending on the previous statistics provided ,in future we are planning to take up same case study run across OpenMP ,Hadoop in distributed environment using different schedulers.

REFERENCES

[1] Parallel Clustering Algorithms: Survey" by Wooyoung KiCSC 8530 Parallel Algorithms Spring 2009
[2] Han, J. and Kamber, M. "Data Mining:Concepts and Techniques", 2000(MorganKaufmann, SanFrancisco, California
[3] Hiral M. Patel "A Comparative Analysis of MapReduce Scheduling Algorithms for Hadoop"IJIERE
[4] www.http://home.deib.polimi.it/matteucc/Clustering /tutorial_html/
[5] R. C. Dubes and A. K. Jain. "*Algorithms for Clustering Data* "Prentice Hall, 1988.
[6] Apache Hadoop. http://hadoop.apache.org
[7] http://hadoop.apache.org/docs/r1.2.1/hdfs_design.
[8] http://hadoop.apache.org/docs/r1.2.1/mapred_ tutorial.html.